Tellus SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Spatial-temporal fractions verification for high-resolution ensemble forecasts

By LE DUC[1,2]*, KAZUO SAITO[1,2] and HIROMU SEKO[1,2], [1]*Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan*; [2]*Meteorological Research Institute, Tsukuba, Japan*

(Manuscript received 13 March 2012; in final form 31 March 2013)

ABSTRACT

Experiments with two ensemble systems of resolutions 10 km (MF10km) and 2 km (MF2km) were designed to examine the value of cloud-resolving ensemble forecast in predicting precipitation on small spatio-temporal scales. Since the verification was performed on short-term precipitation at high resolution, uncertainties from small-scale processes caused the traditional verification methods to be inconsistent with the subjective evaluation. An extended verification method based on the Fractions Skill Score (FSS) was introduced to account for these uncertainties. The main idea is to extend the concept of spatial neighbourhood in FSS to the time and ensemble dimension. The extension was carried out by recognising that even if ensemble forecast is used, small-scale variability still exists in forecasts and influences verification results. In addition to FSS, the neighbourhood concept was also incorporated into reliability diagrams and relative operating characteristics to verify the reliability and resolution of two systems.

The extension of FSS in time dimension demonstrates the important role of temporal scales in short-term precipitation verification at small spatial scales. The extension of FSS in ensemble space is called the ensemble FSS, which is a good representative of FSS for ensemble forecast in comparison with the FSS of ensemble mean. The verification results show that MF2km outperforms MF10km in heavy rain forecasts. In contrast, MF10km was slightly better than MF2km in predicting light rains, suggesting that the horizontal resolution of 2 km is not necessarily enough to completely resolve convective cells.

Keywords: small-scale variability, fractions skill score, intensity-scale diagram, reliability, resolution

1. Introduction

The operational numerical weather prediction (NWP) has been considerably improved by continuous advances in numerical modelling, computer power and data assimilation techniques with new kinds of observations. Quantitative precipitation forecast (QPF) by NWP models was formerly notoriously difficult (Gaudet and Cotton, 1998), but the recent meso-scale models are about to succeed in overcoming this problem for weak to moderate rains. For example, the Japan Meteorological Agency (JMA) has used the meso-scale model (MSM) for operational forecast since 2001. The threat score of this model was about 0.2 in 2001 and increased to about 0.4 in 2011 when verifying on a 20-km grid with a threshold of 5 mm/3 hour (Saito, 2012). However, many difficulties remain in predicting intense rains of correct intensity, location and timing. Application of cloud resolving models to NWP has been started in several forecast centres, and high-resolution data assimilation is essential to further improve short-range forecasts of heavy rain.

One reason for these difficulties lies in the inherent low predictability of local heavy rainfall that occurs under convectively unstable atmospheric conditions (Saito et al., 2011a). To cope with such significant forecast uncertainties of meso-scale severe weather, the use of meso-scale ensemble prediction system (EPS) has also been started in several forecast centres (e.g., Marsigli et al., 2005; Bowler et al., 2008; Wang et al., 2011).

In 2008, an international research project, the World Weather Research Program (WWRP) Beijing 2008 Olympics Research and Development Project (B08RDP), was conducted in conjunction with the Beijing Olympic Games (Duan et al., 2012). Meso-scale ensemble prediction experiments were carried out by six organisations in near real time in order to share their experiences in the development of meso-scale EPSs. Verification for 6-hour rainfall forecasts was performed by Kunii et al. (2011)

*Corresponding author.
Email: leduc@jamstec.go.jp

Citation: Tellus A 2013, **65**, 18171, http://dx.doi.org/10.3402/tellusa.v65i0.18171

using these experiment results, which were interpolated into a common verification domain with a horizontal resolution of 15 km. For all systems, the ensemble means reducing the forecast errors, and the ensemble forecasts clearly improved Brier scores, compared with the control forecasts. However, the detection of intense rains was insufficient for most models, suggesting that the horizontal resolution of 15 km used in B08RDP EPS inter-comparison was inadequate to properly predict strong convective rains. To provide probabilistic information on prevention of natural hazards occurring as a result of meso-scale severe weather events such as heavy rainfall, a higher resolution EPS by cloud-resolving models is required, and the necessity of validation of high-resolution meso-scale EPS is increasing.

Recently, we have carried out experiments using two meso-scale EPSs with the resolutions of 10 and 2 km for 15 d in the summer of 2010. The two ensemble systems were verified to investigate the value of ensemble forecast with increasing resolutions. Considering the merit of meso-scale EPS in predicting local heavy rainfall, we focused on short time (1-hour) rainfall at high resolutions. Intense rains in short time are difficult to predict while sometimes more hazardous in the viewpoint of urban-type disaster prevention. The Fractions Skill Score (FSS) extended to time, and ensemble space was used as the metric for the evaluation. The main issues addressed in those experiments are a) the role of temporal scales in high-resolution verification; b) the unique representative FSS for ensemble forecast; c) the reliability and resolution of ensemble forecast as seen from the neighbourhood view; and d) the outperformance of high-resolution ensemble forecast over low resolution ensemble forecast as measured by the extended FSS.

This article has been organised in the following way. After the introduction, a short overview about the current QPF verification methods for high-resolution forecasts is given in section 2. Section 3 describes the design of the ensemble forecast experiment and verification data used by this study. The first part of section 4 lays out the mathematical foundation of the extended FSS. The mathematical treatment in this section uses the same notations as in Roberts and Lean (2008). Then, the behaviour of FSS when adding the time or ensemble dimension is examined separately. Section 5 deals with the verification results of the two ensemble systems. The last section summarises the main results of this study.

2. Brief review of QPF verification methods

In recent years, a number of new verification methods have been proposed and applied for high- and very-high-resolution precipitation forecasts. At these resolutions, precipitation forecasts become more realistic but at the same time the impact of uncertainties on forecasts due to small-scale processes is more evident. As a consequence, the traditional verification methods do not work properly due to its request of exact matches between forecasts and observations, ignoring small-scale variability.

Almost all methods were proposed to account for spatial mismatches between forecasts and observations. This is because the effect of spatial variability on the traditional scores can be recognised more clearly at high-resolution precipitation forecasts. The simplest cure for this is to calculate these scores in up-scaling grids rather in model grids (Zepeda-Arce et al., 2000). The effect of temporal variability is controlled by performing verification for precipitation of at least 3-hour accumulation.

By focusing on the spatial uncertainty these methods are usually known as the spatial verification methods and were well reviewed in Gilleland et al. (2009). Some promising methods are listed here: Contiguous Rain Area (CRA; Ebert and McBride, 2000), Intensity Scale (Casati et al., 2004), Method for Object-based Diagnostic Evaluation–(MODE; Davis et al., 2006), FSS (Roberts and Lean, 2008), Structure, Amplitude and Location (SAL; Wernli et al., 2008), and Procrustes Shape Analysis (Lack et al., 2010). Ahijevych et al. (2009) carried out idealised and real test cases to gain a basic understanding of behaviour of each method. Some methods have been used routinely in the operational verification systems in several meteorological centres (Mittermaier and Roberts, 2010; Weusthoff et al., 2010).

The spatial verification methods are proposed mainly for deterministic forecast. Some methods introduce the neighbourhood concept for simulating a probabilistic environment and in this way account for spatial variability into deterministic forecast at high resolutions. It is assumed that ensemble forecast can address uncertainties of small-scale processes adequately. However, the finite sample of ensemble members sets a limit on the probability field that an ensemble forecast can represent. Besides, the double penalty problem in high-resolution forecasts due to initial condition and model errors is not reduced even if the number of ensemble members is increased. Thus, the problem of small-scale variability still adheres to high-resolution ensemble forecast. For this reason, the question how to apply the spatial verification methods to ensemble forecast is not counter-intuitive at all.

The up-scaling method is the first spatial verification method proposed for high-resolution deterministic forecast. Marsigli et al. (2008) have attempted to apply the idea of this method into ensemble forecast. They introduced a method called ‘distributional method’ in which comparison between distribution parameters of forecasts and those of observations in a spatial box was performed. The same methodology of the up-scaling method was adopted in

Clark et al. (2011). These authors used relative operating characteristics (ROC) areas over different spatial scales as the verification metrics. The extending of FSS into ensemble forecast was carried out by Schwartz et al. (2010). In their study, the neighbourhood concept was combined with ensemble probabilities yielding neighbourhood ensemble probabilities, which resemble forecast fractions in the original FSS method. Mittermaier (2007) also applied FSS in verification of a lagged ensemble system. The author treated all non-zero probabilities as yes-forecasts, thus transforming the ensemble forecast to a deterministic forecast. Gallus (2010) applied MODE and CRA for every ensemble member. The spreads of rainfall object properties detected by MODE and CRA were used to analyse the spread–skill relationship of the ensemble forecast.

This study performed verification of 1-hour precipitation forecasts using FSS. With such short-term precipitation, the effect of temporal variability on verification results now becomes more significant and should be accounted for in verification. The temporal variability will be addressed by incorporating the time dimension into the fraction concept defined originally in FSS. The ensemble space is also incorporated into fractions to make use of the robustness of ensemble forecast against small scale variability.

3. Design of experiment

Two 11-member ensemble forecast systems MF10km and MF2km, the later nested inside the former with a 6-hour lag, were conducted in the 2010 Baiu season. Both systems used the JMA non-hydrostatic model NHM (Saito et al., 2006; Saito, 2012) as the forecast model. Whereas MF10km applied the modified Kain-Fritsch convective scheme, MF2km did not use convective parameterisation. Other physics processes of the two systems were almost identical to those of the operational MSM and the local forecast model (LFM) of JMA (Hirahara et al., 2011), respectively. The bulk method that predicts mixing ratios of six water species (water vapour, cloud water, rain water, cloud ice, snow and graupel) and number densities of cloud ice were adopted as the cloud microphysical process.

The domains of two systems are illustrated in Fig. 1. The coarse resolution system MF10km had a domain of 361×289 horizontal grid points with 50 vertical levels, forecasted up to 36 hours. For initial conditions, MF10km used the analyses from the JMA non-hydrostatic 4DVAR data assimilation system (Honda and Sawada, 2008). The lateral boundary conditions were interpolated from the forecasts of JMA's high-resolution (TL959L60) global spectral model (GSM). The initial and lateral boundary perturbations were derived from those of JMA's 1-week global EPS (WEP) with a similar normalisation process as described in Saito et al. (2011b; 2012).

The fine resolution system MF2km downscaled MF10km forecasts. This system employed a horizontal resolution of 2 km (800×550 horizontal grid points) with 60 vertical levels. The forecast range is 24 hours. The initial and boundary conditions for each member in MF2km were interpolated directly from the forecasts of the corresponding member in MF10km with a 6-hour lag.

Verification was performed for the precipitation forecasts in July 2010. MF10km started running at 12 UTC

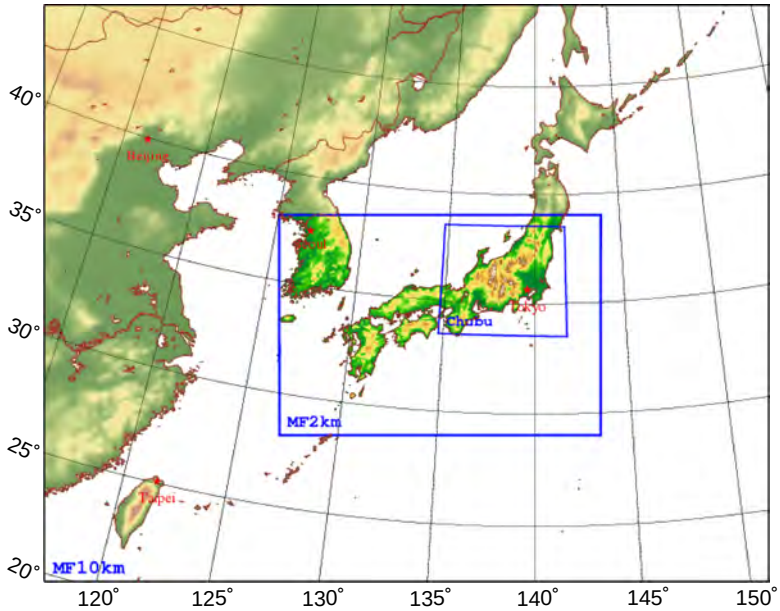

Fig. 1. Forecast domains of MF10km and MF2km. The rectangle inside MF2km domain denotes the verification area.

every day, and MF2km 6 hours later. The forecasts covered the periods of moderate or heavy rainfall events occurring over central Japan. Therefore the dates of forecast dataset did not contain all the dates in this month. Totally, there were 15 forecasts per system in this period, distributed irregularly from 3 July 2010 to 2 August 2010. The rainfall analyses from the JMA's Radar-AMeDAS (R/A) system (Nagata, 2011) were used as references for verification. R/A estimates rainfall every 30 minutes with a horizontal resolution of 1km over Japan area, correcting composite radar echoes by rain gauge observations.

Each system was verified at its native grid to prevent the distortion of rain fields through mapping process (interpolation or filter) when the verification grid was different from the native model grid. However, a common geographic domain was used for two systems in verification. Since R/A provides rainfall observations at 1 km grid spacing, which is finer than the resolutions of two models (10 km and 2 km), this dataset needed to be up-scaled to the native grids of MF10km and MF2km. This was done by taking average over all observation grid points contained inside each model grid cell. Seven spatial scales (20, 60, 100, 140, 180, 220 and 260 km) were chosen for the coarse system, while the fine system used nine smaller ones (04, 12, 20, 60, 100, 140, 180, 220 and 260 km) in computing fractions. There is an overlap between the spatial scales of MF10km and MF2km which was used to compare the performance of the two models.

The verification period were decided based on two factors: temporal scale and spin-up time. Temporal scales involve the extended FSS, which will be discussed in the next section. Since this study concerns the performance of 1-hour precipitation forecasts, the maximum temporal scale was set to 5 hours implying that spatial–temporal boxes admit all 1-hour precipitations valid from 2 hours before to 2 hours after the current time. This means we only examined three temporal scales (1, 3 and 5 hours) in verification. The first 3-hour forecasts from MF2km are considered unreliable due to model spin-up and were discarded in verification. Bringing together these two factors and the objective of comparing performances of two ensemble systems, all verification results were aggregated for whole periods from 6 to 16-hour forecasts by MF2km which were correspondent to 12-hour to 22-hour forecasts by MF10km.

As the first glimpse into the performances of two systems, Fig. 2 shows the accumulated rainfall analysed by R/A and its counterparts forecasted by the control runs of MF10km and MF2km in the whole verification period. The subjective verification over this figure suggests that both control forecasts of MF10km and MF2km predicted well the precipitation amount and location over this period. MF2km provided more detailed distribution of the accumulated rainfall. Rainfalls near the north-west and south-east corners of this figure are probably under-estimated in the R/A precipitation analysis since C-band radar echoes observe upper atmosphere in distant areas and there are no rain gauge observations over the sea.

Verification can be made using the traditional methods as depicted in Fig. 3 with frequency biases (FB). The verification rainfall thresholds vary from light (0.1 mm h^{-1}) to intense (20 mm h^{-1}) rains in Fig. 3. It should be kept in mind that while high thresholds restrict rain events to heavy rains, low thresholds do not only represent light rains but take into account all rain events ranging from light to heavy rains. The FBs point out that MF2km control forecasts under-estimate rain events with low rainfall thresholds and somewhat over-estimate rain events with high thresholds over 20 mm h^{-1}. This implies that MF2km under-estimate light and moderate rain events. In contrast, FBs of MF10km control forecasts are close to unity for light and moderate rains while obviously under-estimate intense rains over 20 mm h^{-1}.

4. Extended FSS

4.1. Mathematical formulation

The FSS results from the normalisation of the Fractions Brier Score (FBS), which in turn is computed from fraction fields. Thus the definition of FSS is based on forecast and observation fractions inside a spatial neighbourhood or window, assumed as a square or circle area centred at each verification pixel. The forecast and observation fractions $M_{(n)}$, $O_{(n)}$ for each window are computed as the ratio between the number of occurrences of the event of interest and the number of grid points in this window. Using a square neighbourhood of size n (also known as a spatial scale), the forecast fraction at a verification pixel (i, j) in a two-dimensional space is defined by

$$M_{(n)}(i,j) = \frac{1}{n^2} \sum_{ii=1}^{n} \sum_{jj=1}^{n} I_M(ii,jj) \qquad (1)$$

where I_M has a binary value depending on a yes-forecast (1) or no-forecast (0) event. The index ii, jj run over all verification pixels inside the neighbourhood. The mathematical formula for $O_{(n)}$ has similar form with I_M replaced by I_O.

This concept of fractions can be extended seamlessly into a three-dimensional space by adding another summation symbol in the right-hand side of eq. (1). Instead of a neighbour area in space, rather a neighbourhood should be

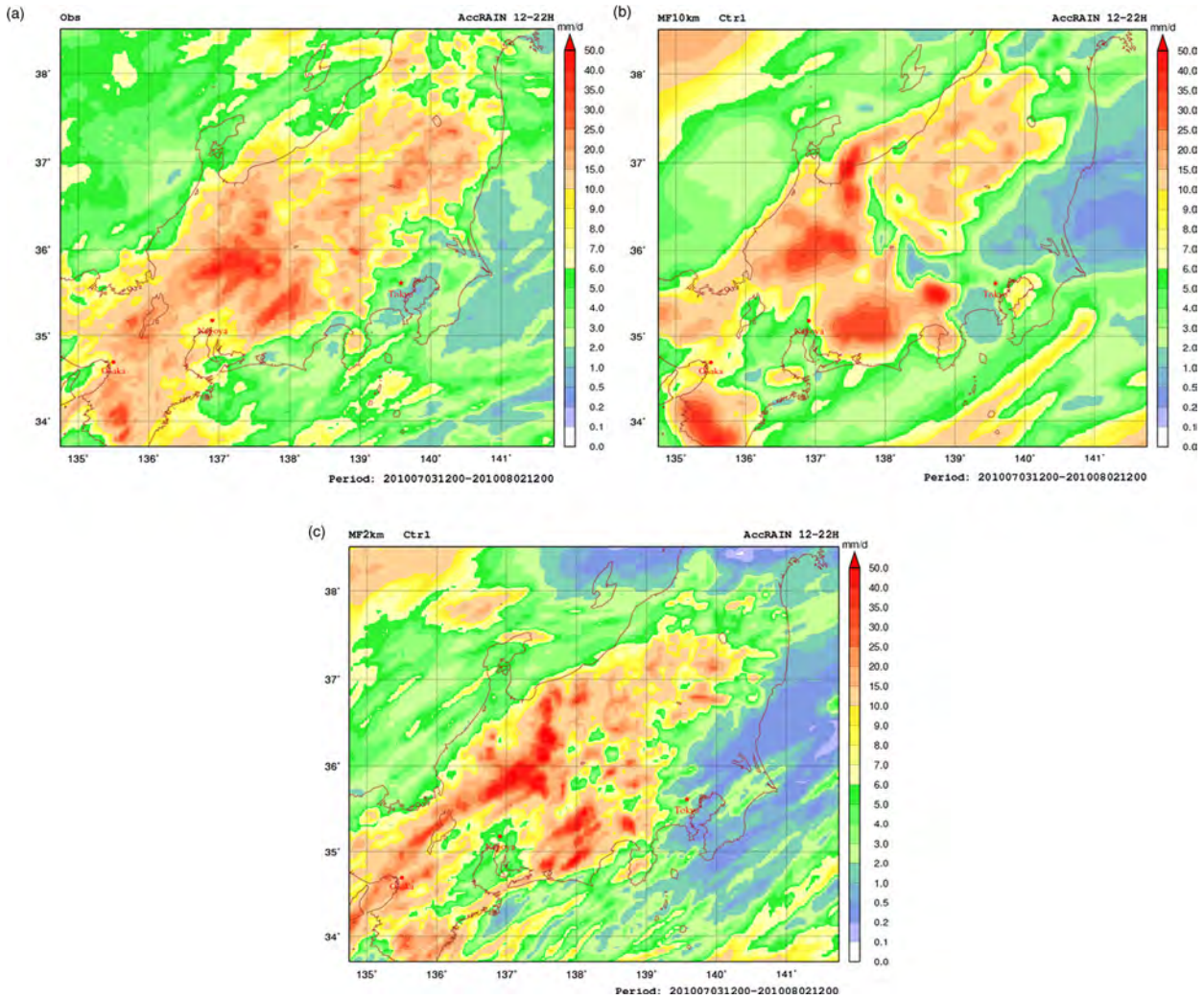

Fig. 2. Rainfall analysis by R/A (upper-left) and corresponding control forecasts by MF10km (upper-right) and MF2km (lower-left) in the verification period over central Japan. Rainfalls were accumulated between 12-hour and 22-hour forecast ranges and the average rain rates (unit mm d^{-1}) are shown in the plot.

understood as a spatial–temporal box, hence eq. (1) becomes

$$M_{(nm)}(i,j,k) = \frac{1}{n^2 * m} \sum_{ii=1}^{n} \sum_{jj=1}^{m} \sum_{kk=1}^{m} I_M(ii,jj,kk) \tag{2}$$

Here, the new index k and kk stand for the new dimension, namely the time dimension. To distinguish from the spatial scale n, the temporal scale is denoted as m.

The uncertainties of small-scale processes in space and time can be sampled using such spatial–temporal box. However, this strategy does not sample well enough other sources of uncertainty, for example, initial condition deficiencies or model errors. The fact that ensemble forecast has been used to quantify this kind of uncertainty suggests that the concept of fractions can apply for ensemble forecast by incorporating the ensemble dimension into a neighbourhood. The ensemble dimension corresponds to the space where each member from an ensemble forecast is considered as a possible realisation of the true state. With the ensemble dimension added, eq. (1) leads to

$$M_{(nmp)}(i,j,k,l) = \frac{1}{n^2 * m * p} \sum_{ii=1}^{n} \sum_{jj=1}^{n} \sum_{kk=1}^{m} \sum_{ll=1}^{p} \times I_M(ii,jj,kk,ll) \tag{3}$$

where l and ll are the index of the ensemble dimension, and p the number of ensemble members taking into account. Equation (3) shows a forecast fraction defined in a four-dimensional space and dependent on three scale parameters n, m and p.

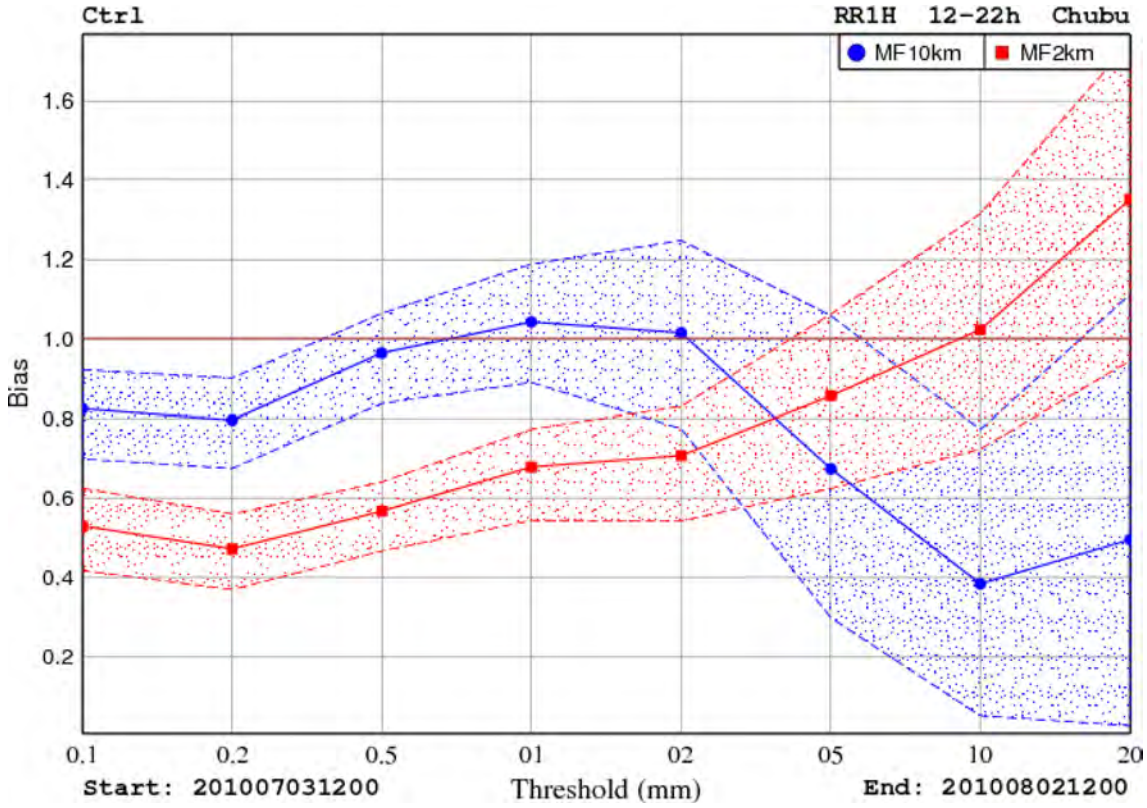

Fig. 3. Frequency bias of hourly precipitation forecasts from MF10km and MF2km control runs in July 2010. The shaded areas are the 95% confidence intervals.

Note that whereas the space dimension and the time dimension have a sense of order, the ensemble dimension does not. For a given number of members p, there exists a vast combination of the neighbour members around a member l, which is different from the unique set of the neighbour points at any grid point for a given spatial or temporal scale. In this case, the fractions should be averaged over all possible combinations of neighbour members. If the ensemble forecast has N members, the number of p-combinations of ensemble members is simply the binomial coefficient $C(N,p)$. Equation (3) should be rewritten, restricting here to the ensemble dimension for simplicity

$$M_{(p)} = \frac{1}{C(N,p) * p} \sum_{ip=1}^{C(N,p)} \sum_{ll=1}^{p} I_M(ip, ll) \qquad (4)$$

with the index l discarded due to the independence of $M_{(p)}$ with any specific ensemble member l. The number of p-combinations that contain the ensemble member l is $C(N-1,\ p-1)$, which can be easily verified if a p-combination from N elements with an element A inside can be considered as a combination of A with a $(p-1)$-combination from $(N-1)$ remaining elements. That means the number of occurrences of I_M in eq. (4) is similar for all ensemble members and equal to $C(N-1, p-1)$. Hence, eq. (4) reduces to

$$\begin{aligned} M_{(p)} &= \frac{1}{C(N,p) * p} \sum_{ll=1}^{N} C(N-1,\ p-1) I_M(ll) \\ &= \frac{C(N-1,\ p-1)}{C(N,p) * p} \sum_{ll=1}^{N} I_M(ll) = \frac{1}{N} \sum_{ll=1}^{N} I_M(ll) \end{aligned} \qquad (5)$$

Here, we obtain an interesting result that fractions based on averaging over all possible combinations of subsets of p members from an N-member ensemble is identical to fractions based on all N members. This reduces the computational cost considerably, since the summation in eq. (5) is reduced by a factor of $C(N,p)*p/N$ in comparison with eq. (4). Equation (3) becomes

$$\begin{aligned} M_{(nmN)}(i,j,k) = &\frac{1}{n^2 * m * N} \sum_{ii=1}^{n} \sum_{jj=1}^{n} \sum_{kk=1}^{m} \sum_{ll=1}^{N} \\ &\times I_M(ii, jj, kk, ll) \end{aligned} \qquad (6)$$

Now, the FBS can be defined the same as the one in Roberts and Lean (2008) (these authors called it mean

square error (MSE) in their article) by averaging the differences between forecast and observation fractions over all verification pixels *(i,j and k)* in the verification domain:

$$FBS_{(nmN)} = \frac{1}{N_x N_y N_t} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{N_t} [M_{(nmN)}(i,j,k) - O_{(nm)}(i,j,k)]^2 \quad (7)$$

Here, N_x and N_y are the number of verification pixels in the x and y axis, respectively; N_t is the number of time slices.

FSS has the same form as proposed by Roberts and Lean (2008) and is reproduced here for the sake of completeness

$$FSS_{(nmN)} = \frac{FBS_{(nmN)} - FBS_{(nmN)ref}}{FBS_{(nmN)perfect} - FBS_{(nmN)ref}} = 1 - \frac{FBS_{(nmN)}}{FBS_{(nmN)ref}} \quad (8)$$

where the zeros value of the perfect FBS has been applied implicitly and the reference FBS in a four-dimensional space has the following form

$$FBS_{(nmN)ref} = \frac{1}{N_x N_y N_t} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{N_t} [M^2_{(nmN)}(i,j,k) + O^2_{(nm)}(i,j,k)] \quad (9)$$

4.2. FSS with the time dimension included

Using FSS as a metric, performance of a deterministic model is usually summarised in an intensity-scale diagram (Ebert, 2008). Each square in this diagram is coloured after its FSS value which varies with spatial scale and rainfall intensity. These two parameters are expressed in the horizontal and vertical axes, respectively. Normally, FSS increases with spatial scale and tends to 1 asymptotically if forecasts are unbiased. When rainfall intensity increases, FSS usually decreases indicating that performance gets lower when forecast objectives shift from light to intense rains. Using this diagram, users can easily identify good forecast areas which consist of skilful spatial scales at certain rainfall thresholds.

The incorporation of the time dimension into FSS requires a new form for intensity-scale diagrams since the impact of temporal scales on FSS should be addressed. In the simplest way, for each temporal scale, an intensity-scale diagram as described above can be provided, and the impact of temporal scales on FSS can be inferred by comparing two intensity-scale diagrams. Or to simplify the comparison, a diagram with the same form as the intensity-scale diagram can be created, showing only FSS differences between two distinct temporal scales. This visualisation strategy is clearly not a good solution since a lot of diagrams need to be produced, and the relationship between spatial and temporal scales via FSS is not easy to explore.

In this study, we proposed a modified intensity-scale diagram, which comprises spatial and temporal scales together with rainfall intensities. This new intensity-scale diagram is illustrated in Fig. 4 using the control runs from MF10km and MF2km as forecasts. In the new diagram, the spatial scale and intensity axes are kept as in the original form. For each intensity value, a horizontal temporal scale axis will be embedded, resulting in a spatial–temporal sub-diagram inside the overall intensity-scale diagram. Since the number of temporal scales for hourly precipitation is limited to three (equivalent to a maximum 5-hour temporal scale), the horizontal length of the modified diagram is not elongated and is reasonable to follow.

As expected, FSS increases with increasing of spatial or temporal scales in Fig. 4, which means that the performance in forecasting short-term precipitation will be underestimated if temporal lag is not accounted for. This clearly demonstrates the importance of temporal uncertainties when short-term precipitation forecasts are verified in context of high-resolution forecasts.

Further investigation can identify FSS-constant lines with an approximated slope of –10 km/1 hour in each spatial–temporal plane for both control forecasts in Fig. 4. These constant curves show that the FSS values at small spatial and long temporal scales are equal to the ones at large spatial and short temporal scales, for example, FSSs at 20 km, 5-hour scales and 60 km, 1-hour scales are similar. This fact suggests that MF10km and MF2km forecasts may have an error of 10 km/h in estimating propagation speed of rainfall systems. However, the answer to the question that whether the forecasts had early or late biases cannot be determined by the fact that neighbourhoods are symmetric around any grid point. Another implication from this result is that the slope of the FSS-constant lines may be affected by the spatiotemporal scales of meso-scale phenomena (e.g., 10 km and 1 hour for cumulonimbus, and a few tens of kilometres and hours for meso-scale convective systems).

There exists a distinct change of FSSs between the 2 and 5 mm h^{-1} rainfall thresholds in the intensity-scale diagram of MF10km control forecasts, whereas such large change of FSSs does not appear in the one of MF2km control forecasts where the FSSs vary smoothly from threshold to threshold. This shows in an illustrative way that MF10km control forecasts could not capture well convective intense rains, which can attribute to the limit of the Kain–Fritsch convective parameterisation scheme. It is known that JMA's operational MSM has a gap of QPF performance between 10 and 20 mm/3 hours. As a convection-permitting model,

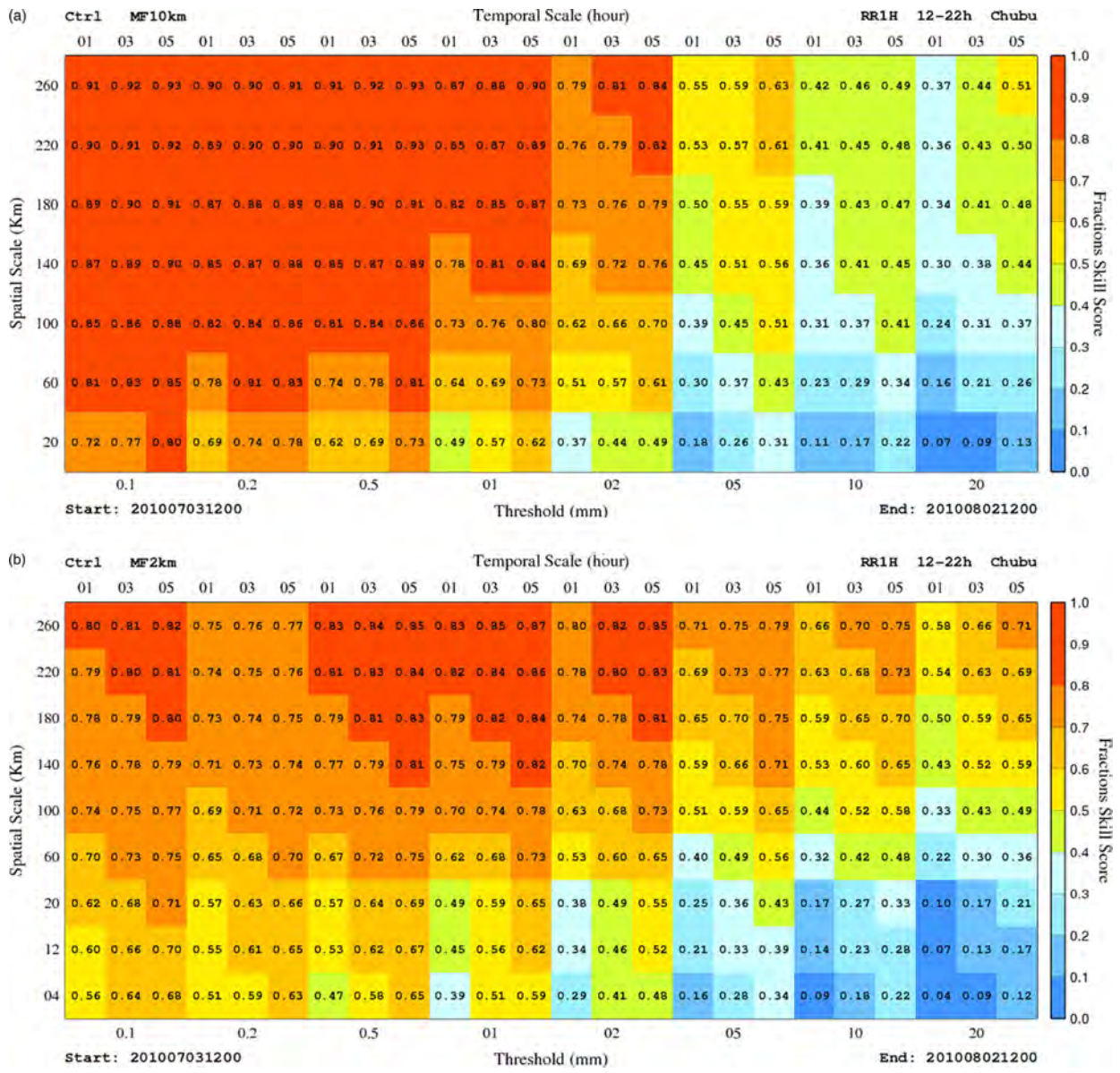

Fig. 4. Intensity-scale diagrams with temporal scales incorporated from the control forecasts of MF10km (top) and MF2km (bottom).

MF2km control could represent convective precipitation more properly and results in a better forecast with respect to heavy rainfall.

4.3. *FSS with the ensemble dimension included*

This subsection aims to examine the behaviour of FSSs in ensemble space and find an appropriate FSS characterising for ensemble forecast. Since spatial and temporal scales play similar roles in this problem, we will handle FSS in the absence of temporal scales. FSSs depending both on spatial scales and temporal scales will be addressed in section 5.

Fractions in the ensemble space can be determined based on two operators, namely the ensemble mean operator EM and the threshold operator TC. If all ensemble members are denoted by Ω, the probability of a yes-forecast event at a particular space and time can be calculated as TC[EM(Ω)] or EM[TC(Ω)]. The only difference is in the order of operators. We obtain a binary value in the first definition and a fraction in the second one. More specifically, in the first definition, the average was taken over all rain fields before applying a threshold for the averaged field. In the second one, a threshold was applied for the rain field of each member before taking average over all resulting

yes–no masks. Using these fractions in calculating FSS, the FSS of ensemble mean in the first case and the ensemble FSS in the second case can be determined.

An idealised experiment which kept the same configuration as that used by Roberts and Lean (2008) was conducted to investigate the behaviour of the FSS of ensemble mean and the ensemble FSS. A 1-pixel wide observation rain-band and its forecasted counterpart, which was exactly the observation rain-band but shifted 11 pixels, were given in a domain of 100×100 pixels. Thus, a forecast with the displacement error of 11 pixels was supposed. To create an 11-member ensemble forecast, 10 additional forecasts were issued by shifting the given forecast forward or backward around its location, 1–5 pixels. The original forecast was considered as the control forecast. The ensemble mean was derived from the 11 members, and its FSS curve against spatial scales is plotted in the same chart with the ensemble FSS (Fig. 5). Here, the threshold was selected low enough that no precipitation area was disregarded in the ensemble mean. The deterministic FSSs of all member forecasts were also plotted for reference.

The FSS curve from each ensemble member in Fig. 5 represents what was found in Roberts and Lean (2008) saying that FSS values are equal to zeros for all spatial scales less than or equal to displacement errors. Since the ensemble FSS and the FSS of ensemble mean were computed using all members, it is quite understandable that these two curves have the zero values only when the spatial scales are smaller than the minimum displacement error of all members. This means that even when a control forecast shows an unskilful forecast via a FSS value of zero, these two FSS values may differ from zero, showing that the ensemble system owns a certain skill in which good forecasts occur in some members different from the control. However, whereas the FSS of ensemble mean indicates a biased forecast where the FSS values are always smaller than 0.2, the ensemble FSS is quite close to other ensemble member FSSs which tend to one asymptotically, indicating an unbiased forecast. This biased forecast results from an 11-pixel wide rain-band forecasted by the ensemble mean instead of 1-pixel wide rain-bands by other members.

The behaviour of FSSs in real cases with MF10km and MF2km forecasts is shown in Fig. 6. This figure presented different FSS curves under various rainfall thresholds. Again the FSSs of ensemble mean have similar behaviour as one in the idealised case with respect to intense rains (the rainfall threshold of 20 mm h^{-1} in Fig. 6). However, this does not hold when the rainfall threshold decreases. At the rainfall threshold of 2 mm h^{-1}, the change of the FSS of ensemble mean with spatial scale is analogous with those of ensemble member forecasts. The most interesting thing appears at the rainfall threshold of 0.2 mm h^{-1} when the ensemble means show as the best forecasts in term of FSS in comparison with the ensemble member forecasts.

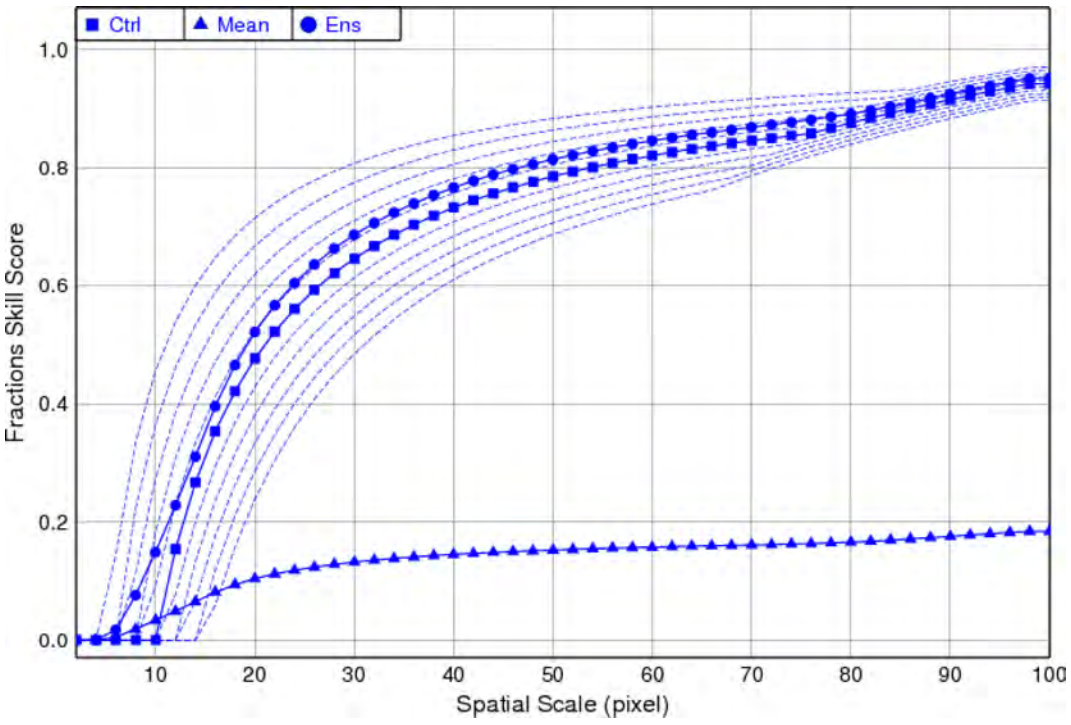

Fig. 5. Ensemble FSS (circle symbol), control FSS (square symbol), ensemble mean FSS (triangular symbol) and FSSs from other ensemble members (dash lines) against spatial scales in the idealised experiment.

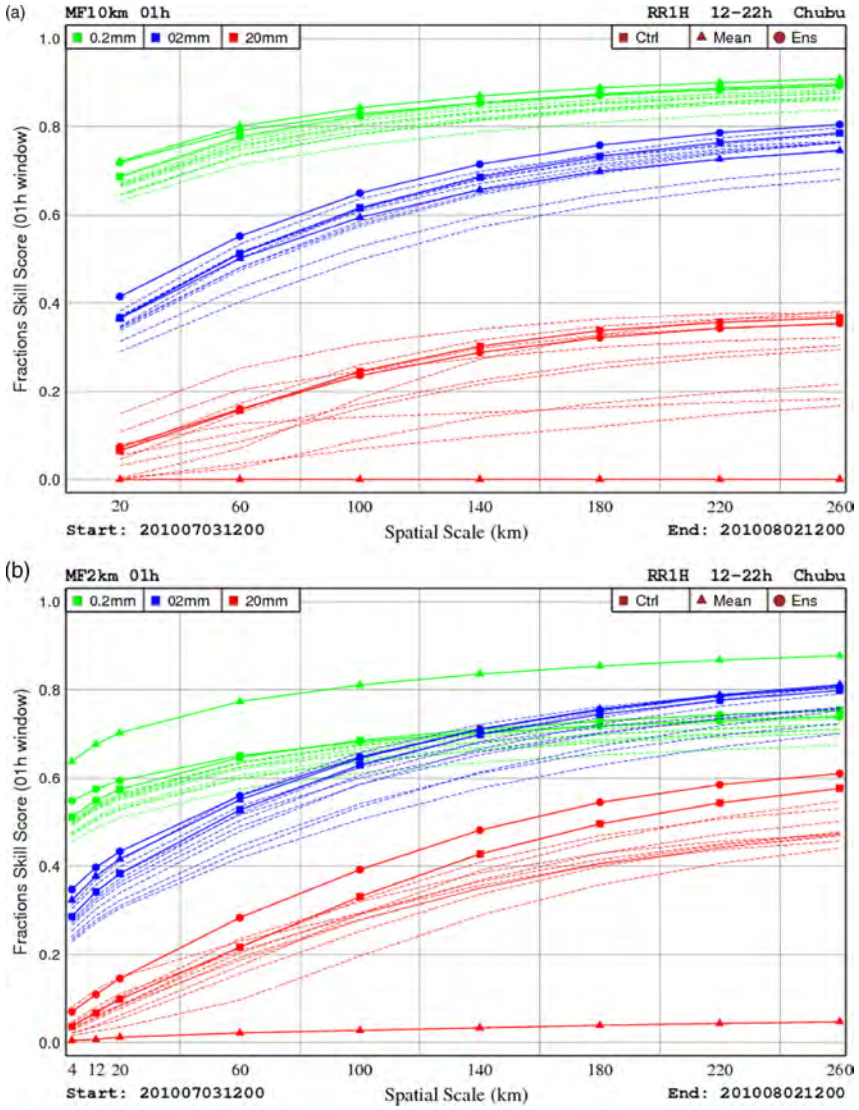

Fig. 6. Same as Fig. 5 but in real cases with MF10km (top) and MF2km (bottom) forecasts in July 2010. The green, blue and red colours represent the rainfall threshold of 0.2, 2, and 20 mm h^{-1}, respectively.

To explain these results, we use the fact that ensemble means tend to smear out rain fields. As a consequence, at low rainfall thresholds, an ensemble mean produces the number of yes-events more than any other ensemble members. Thus, the precipitation area forecasted by the ensemble mean tends to be the superposition of all precipitation areas forecasted by each member.[1] Clearly, when all ensemble forecasts underestimate precipitation areas and the ensemble mean does not overestimate precipitation areas, the ensemble mean will outperform all members in prediction of precipitation areas. This is illustrated in Fig. 6, especially in the case of MF2km. The figure also indicates that precipitation areas are predicted worse in MF2km than in MF10km. In contrast, when all ensemble members overestimate precipitation areas, the ensemble mean will give the worst forecast since the precipitation area forecasted by the ensemble mean is the superposition of all precipitation areas forecasted by all members. Since both MF10km and MF2km underestimate precipitation areas, this case is not observed but can be easily verified in an idealised experiment.

Ensemble means not only smear out but also smooth out rain fields. That means at high rainfall thresholds, an ensemble mean produces an excess of no-events. Therefore,

[1]Here, precipitation areas are identified with the areas covered by yes-events at low rainfall thresholds.

the resulting forecasts score their FSS values smaller than those of other ensemble member forecasts (the 20 mm h^{-1} lines in Fig. 6). This fact in conjunction with the previous analysis suggests that we should not use the FSS of ensemble mean to validate an ensemble forecast since it does not properly reflect the actual performance of ensemble forecast.

The ensemble FSS exhibits the same behaviour as the FSS of ensemble mean with respect to low rainfall thresholds when the FSS values are usually greater than those of ensemble members. For the high-resolution forecasts (MF2km) such behaviour is even observed when the rainfall threshold increases. Moreover, the differences between the ensemble FSS values and the best FSS values of member forecasts are more distinct when rainfall thresholds become higher. At such thresholds, the ensemble FSS curves of MF10km no longer lie above other FSS curves of ensemble members and their behaviour is analogous to the one in the idealised experiment.

The behaviour of the ensemble FSS can be grasped by limiting ensemble forecasts to the simplest case with two members only. Since neighbourhood is the essential element in the definition of FSS, we restrict the calculation of FSS to a neighbourhood of 10×10 verification pixels. Assume that at a predefined threshold the observational frequency is 10/100. If both forecasts underestimate/overestimate this frequency, the ensemble forecast frequency will also be underestimated/overestimated in comparison to the observation frequency. For example, if the frequency for member 1 is 14/100 and for member 2 is 20/100, the ensemble forecast frequency will be $(14+20)/200 = 17/100$. In both cases, the ensemble FSS curves will run between the FSS curves of member 1 and member 2. However, when a member underestimates and another overestimates the rainfall probability, the situation will change drastically. Now, keep the forecasted probability of member 1 and assume the one given by member two is 6/100. The resultant ensemble probability becomes $(14+6)/200 = 10/100$, which is identical to the observational frequency. Thus, we have a perfect forecast in term of FSS where the FSS value is equal to 1, although in this case both underestimated and overestimated forecasts by member 1 and member 2, respectively, have FSS values smaller than 1.

This simple example explains why we see the different behaviours of the ensemble FSS curves in MF10km and MF2km. At high rainfall thresholds, all ensemble members of MF10km underestimate observation fractions, and so does the ensemble envelope. In this case, the ensemble FSS curve is indistinguishable from the FSS curves of all ensemble members. With increasing resolution, some members in MF2km could reproduce heavy rainfall events with precipitation fractions larger than that of observation. Here, the situation is analogous to the example above when some members overestimate while others underestimate observation fractions. Hence, the ensemble FSS curve lies above all the FSS curves of ensemble members.

At low rainfall thresholds, in addition to the mechanism described above, we should note that the fraction field or probability field produced by the ensemble envelope also covers similar area as the precipitation area forecasted by the ensemble mean. The foregoing remarks about the FSS of ensemble mean still hold for the ensemble FSS. The ensemble FSSs tend to be higher than all FSSs of ensemble members. However, the fact that we use fractional instead of binary probabilistic fields in calculation of FSS causes the differences between the ensemble FSS and the FSSs of ensemble members to be less distinct as in the case of the FSS of ensemble mean. Compared with the FSS of ensemble mean, as a FSS metric characterising for an ensemble forecast, the ensemble FSS should be selected.

5. Verification results

5.1. Traditional verification

Before performing verification on MF10km and MF2km forecasts using the ensemble FSS as the metric, the forecast performance as measured by the traditional scores is investigated. This traditional verification was conducted in terms of reliability and resolution using reliability diagrams and ROC at the grid scale of each system. Note that statistical scores were calculated on the different grid resolutions (here 10 and 2 km for MF10km and MF2km, respectively), and we should be careful when making any comparison based on these scores.

Figure 7 shows the reliabilities of MF10km and MF2km through reliability diagrams with three rainfall thresholds. Since each ensemble system had 11 members, the forecast probabilities were divided into 11 bins, namely 0–0.05, 0.05–0.15, ..., 0.95–1. The first diagram for the 0.1 mm h^{-1} threshold indicates that MF10km forecasts are reliable for light rains, whereas MF2km forecasts are under-forecasting in the regime less than 70%. In case of moderate rains (1 mm h^{-1}), both systems exhibit over-forecasting in the regime by more than 30%. MF2km owns certain skill, even though observation frequencies are smaller than the predicted ones in the regime by more than 30%. The tendency of over-forecasting is more evident when heavy rains are considered (5 mm h^{-1}). Both reliability curves diverge from the perfect reliability line in the regime by more than 20%, revealing that performance is lost with respect to heavy rains in the conventional statistics.

The resolutions are shown in Fig. 8 with ROC diagrams. Ten forecast probability thresholds ranging from 0.1 to

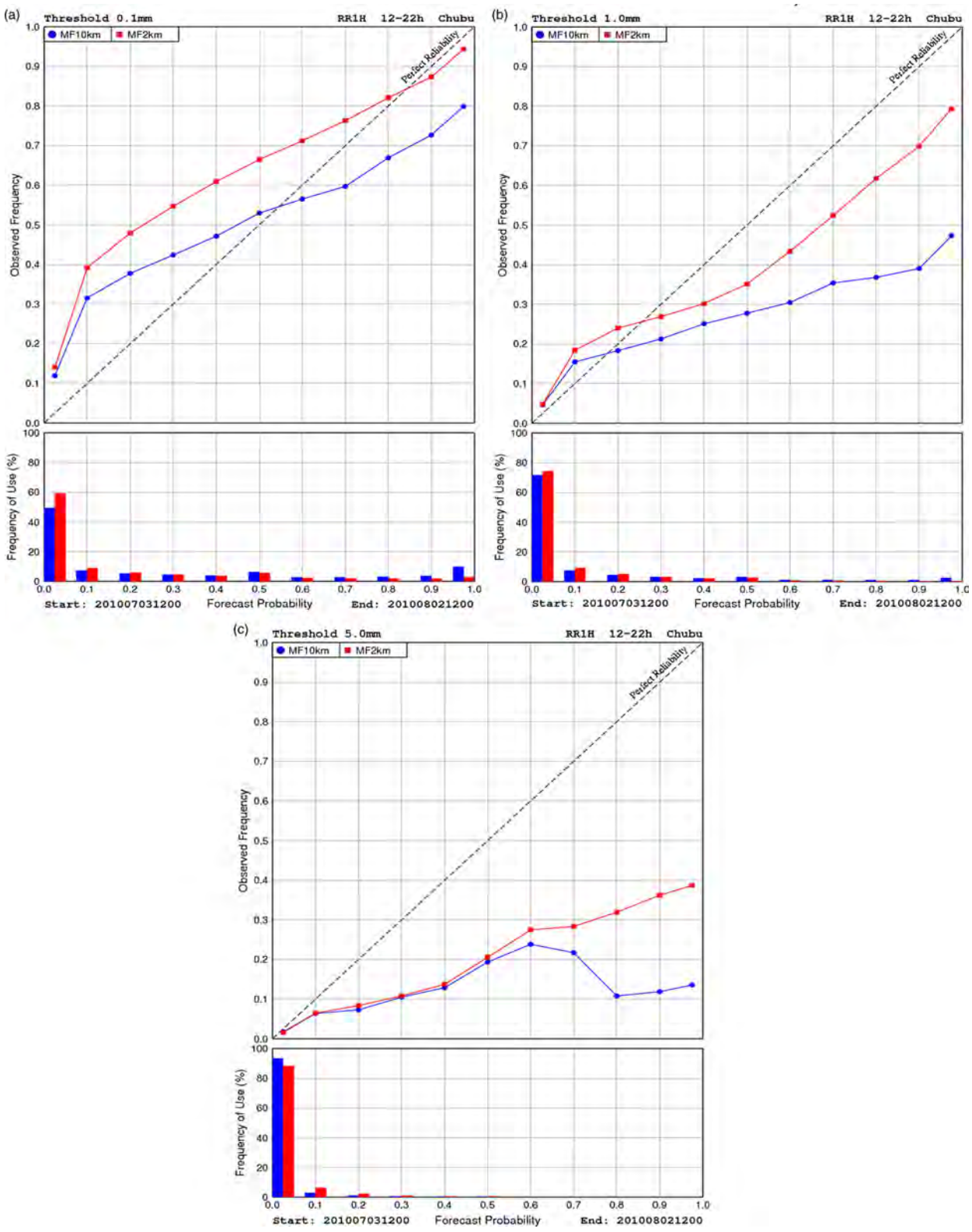

Fig. 7. Reliability diagrams at grid point scales of hourly precipitation forecasts from MF10km and MF2km in July 2010 with the rainfall threshold of 0.1 mm h^{-1} (upper left), 1.0 mm h^{-1} (upper right) and 5 mm h^{-1} (lower left). The sharpness diagram is shown below each reliability diagram.

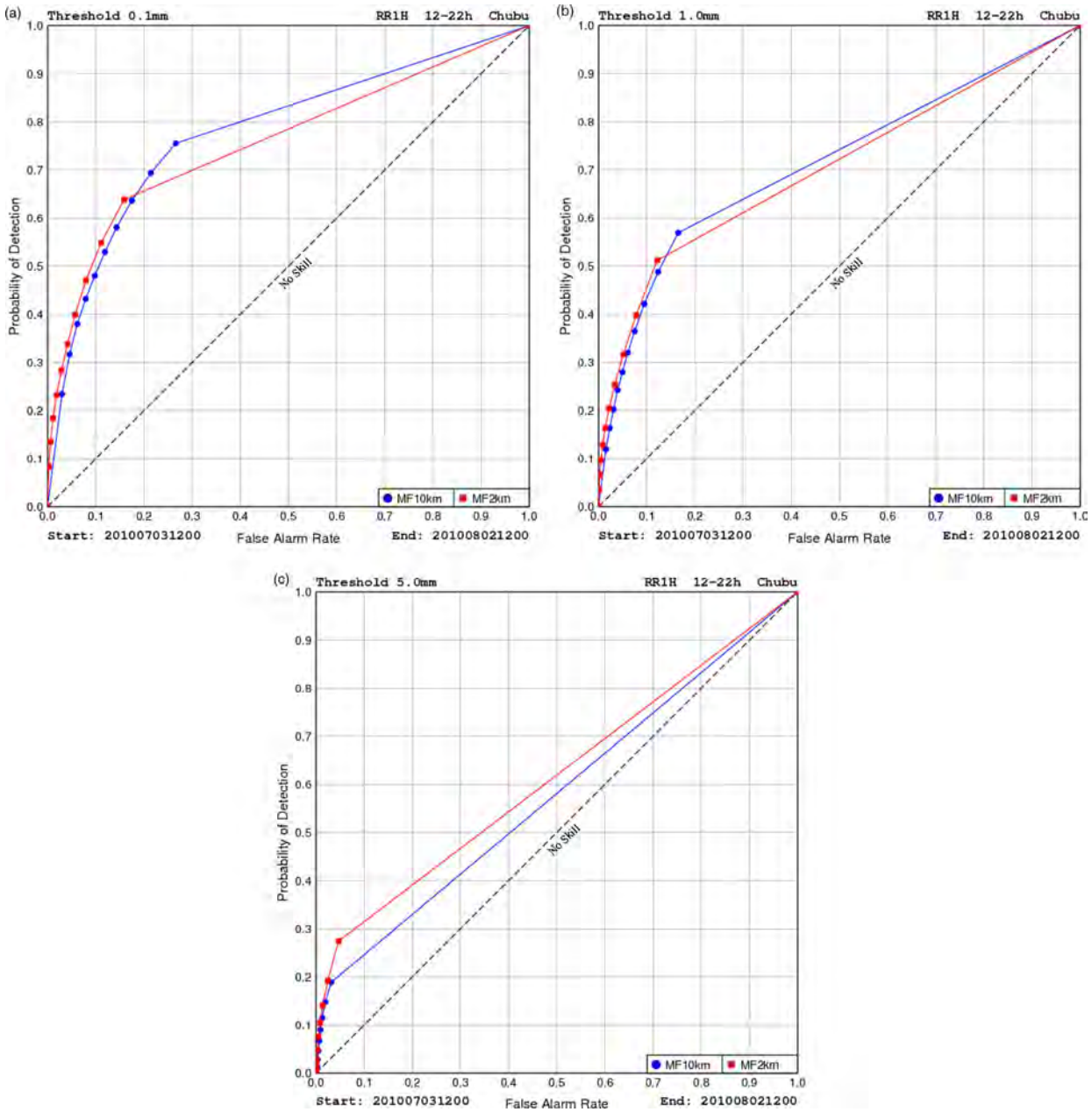

Fig. 8. ROC diagrams at grid point scales of hourly precipitation forecasts from MF10km and MF2km in July 2010 with the rainfall threshold of 0.1 mm h^{-1} (upper left), 1.0 mm h^{-1} (upper right) and 5 mm h^{-1} (lower left).

1.0 were used to produce the ROC curves. It can be seen that the ROC areas decrease when moving from low to high rainfall thresholds. The difference of the ROC areas between MF10km and MF2km is not significant for moderate (1 mm h^{-1}) rain. In case of light (0.1 mm h^{-1}) rain, MF2km is better than MF10km in term of resolution. However, this reverses when considering intense rain, with MF2km discriminating more heavy rain events than MF10km.

The Brier Skill Scores (BSS) that summarise the skills of MF10km and MF2km in both reliability and resolution are given in Fig. 9. The BSS curves indicate that two systems have no skills at medium and high rainfall thresholds. This can be attributed to the over-forecasting[2] at these rainfall

[2]The terminology may confuse the readers. In fact, MF10km under-predicts heavy rainfall events as shown in frequency bias (Fig. 3) but seemingly over-forecasts in Fig. 7. Furthermore, MF10km forecasts for a high probability of heavy rainfall are very rare, and most of them are issued as false alarms. See discussion on Figs. 10 and 11.

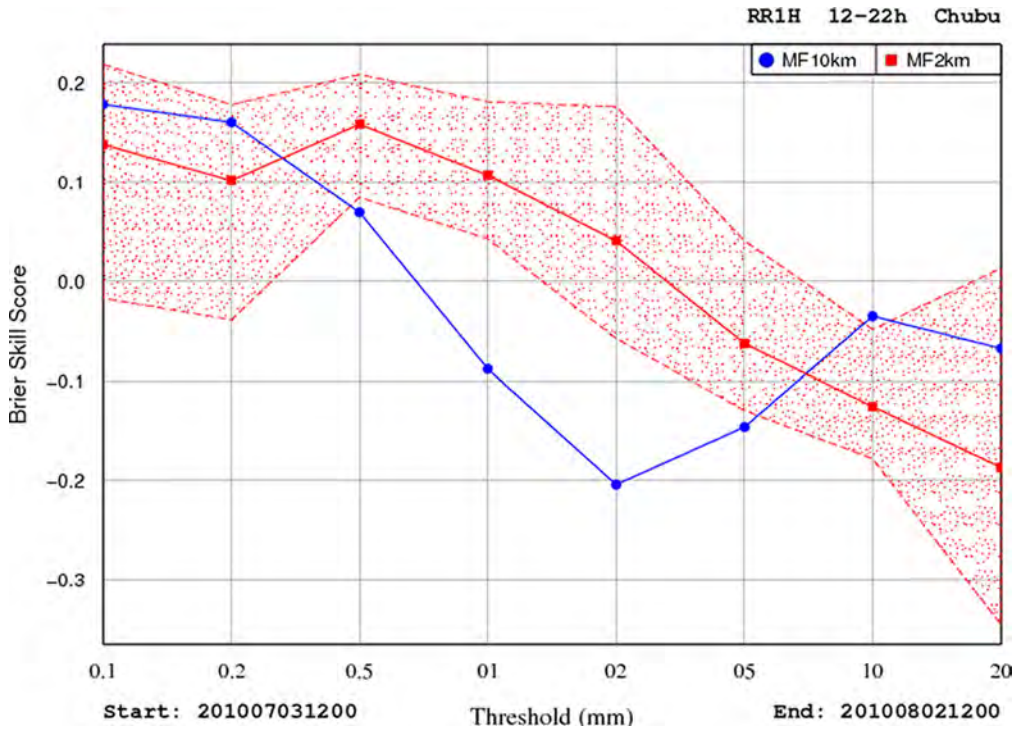

Fig. 9. BSS of hourly precipitation forecasts from MF10km and MF2km in July 2010. The shaded area is the 95% confidence interval for the differences centred to the BSS curve of MF2km.

thresholds. BSSs of MF2km are even worse than those of MF10km for very intense rains of 10–20 mm h^{-1}. The skills are only assessed at the low rainfall thresholds, where both systems have similar performance. This objective result clearly differs from the subjective evaluation, as well as the accumulated rainfall distributions depicted in Fig. 2.

5.2. *Verification with neighbourhoods*

Based on the idea of using neighbourhoods to account for uncertainties in high-resolution forecasts, the reliabilities of MF10km and MF2km are examined again with the incorporation of neighbourhoods into reliability diagrams. When considering at the same spatial and temporal scales, this enables a comparison of the performances of MF10km and MF2km, which is clearly an advantage over the traditional approach in the preceding subsection. Forecast probabilities in reliability diagrams were not considered as the ones computed from fractions of yes-forecasts at verification pixels but instead were identified with forecast fractions in subsection 4.1. To keep consistency with the treatment of observation frequencies in FSS, for each forecast probability, the binary value of the corresponding observation frequency in the traditional reliability diagrams allows varying between 0 and 1, which is identical to an observation fraction in the terminology of FSS. This implies that if forecast probabilities are viewed in a scale, observation frequencies should be done the same way instead of continuing to be viewed at grid scales. At grid scales, this reliability diagram becomes the traditional reliability diagram.

Figures 10 and 11 demonstrate the resulting reliability diagrams with a specific temporal scale of 5-hour (2-hour lag) and five spatial scales (04, 12, 20, 60, 100 km) or with a specific spatial scale of 60 km and three temporal scales ranging from 1 to 5 hours. Although Fig. 10 contains five spatial scales, only three are plotted for each system (04, 12 and 20 km in case of MF02km and 20, 60 and 100 km in the case of MF10km). Since forecasts with high fractions at large scales are rare, especially when combining with heavy rains, the sample sizes at the bins of high forecast probabilities like 0.85–0.95 or 0.95–1 may be almost zeros. In these situations the reliability curves were just plotted for the bins with non-zero sample sizes. This explains why some reliability curves do not go through all bins in Figs. 10 and 11. Note that some points with small sample sizes were still plotted, which lack statistical significance due to under-sampling and should not be used in interpretation, for example, points at the bins of high forecast probabilities for the rainfall threshold of 5 mm h^{-1} in Figs. 10 and 11.

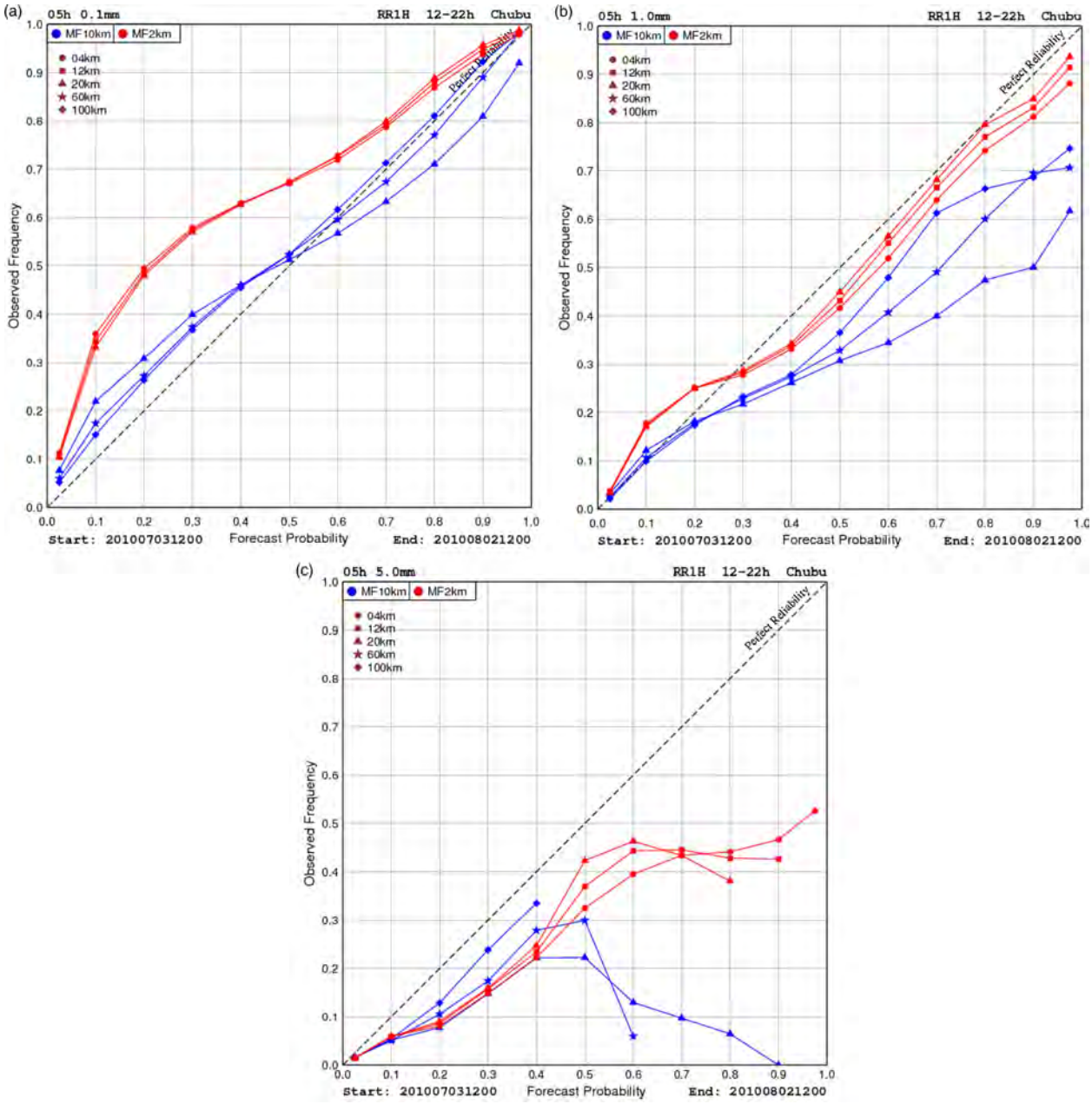

Fig. 10. Same as Fig. 7 but for temporal scale of 5 hours and spatial scales of 4, 12, 20, 60 and 100 km.

The new reliability diagrams clearly show that forecasts are more reliable when spatial or temporal scales increase. The only exception occurs with MF2km forecasts at low rainfall thresholds when the reliability curves tend to go away the perfect reliability line at the bins of high forecast probabilities. These reliability curves also display the under-forecasting bias of MF2km. When this bias is removed, it will be seen that forecast reliability increases with increasing spatial or temporal scale. The same conclusions in the verification with traditional reliability diagrams can be deduced here: at low rainfall thresholds, MF10km exhibits a good reliability whereas MF2km is under-forecasting; if rainfall thresholds are higher than the medium threshold, two systems are over-forecasting. However, at these thresholds MF2km forecasts are more reliable than these of MF10km distinctly when evaluating at the same spatial and temporal scale.

With the success of combining the traditional reliability diagram with the neighbourhood idea in examining reliability of ensemble forecasts, similar methodology was applied for the traditional ROC diagram. Yes-forecast events were defined in the same way as in the traditional

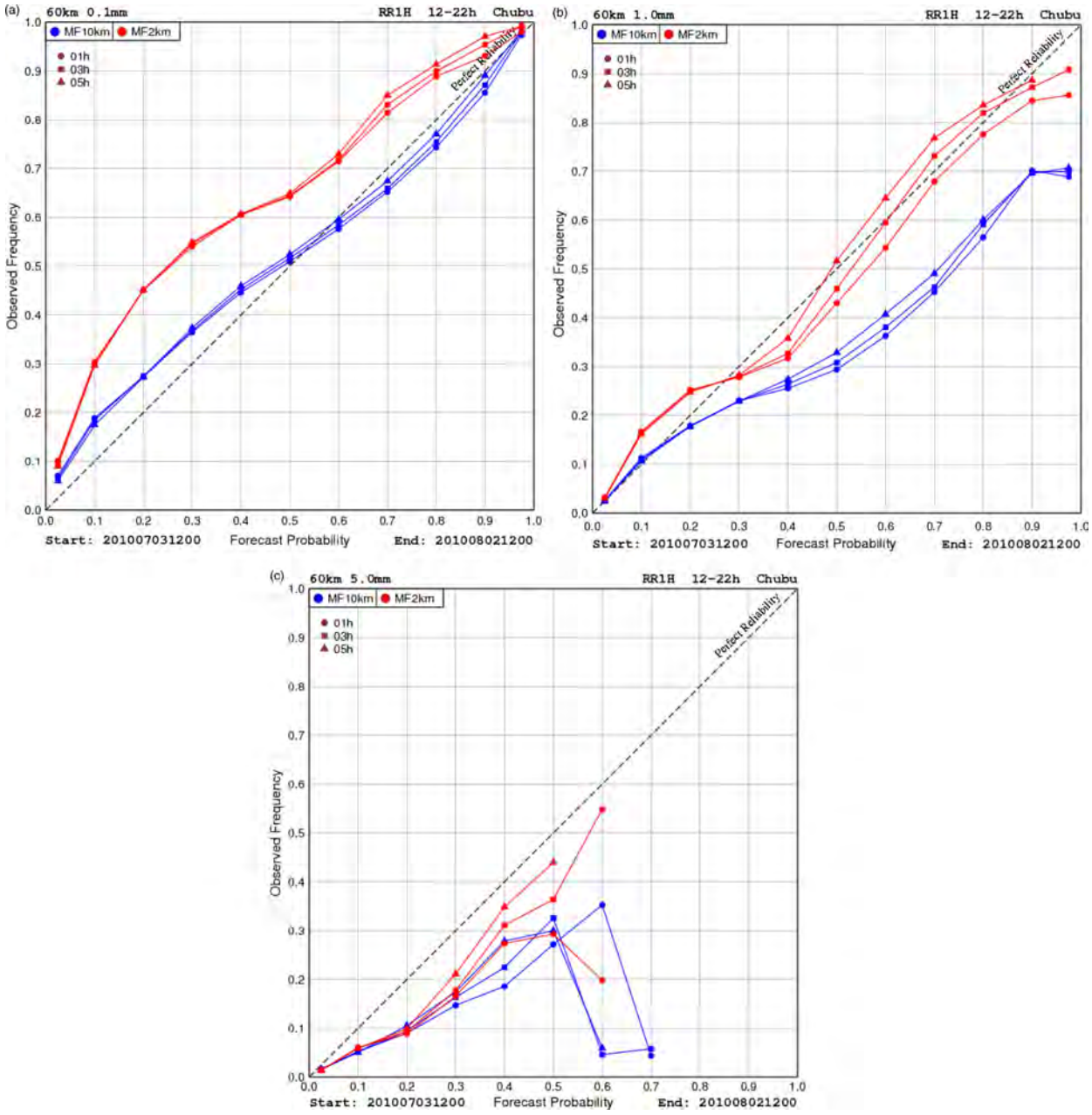

Fig. 11. Same as Fig. 7 but for spatial scale of 60 km and temporal scales from 1 to 5 hours.

ROC with ensemble probabilities replaced by forecast fractions, that is, a yes-forecast event is registered when a forecast fraction is higher than a given threshold. Definition for yes-observation events is trivial as in the traditional ROC when yes-observation events are considered at grid points. At grid scales, the new ROC diagram reduces to the traditional ROC diagram.

The resulting ROC curves are displayed in Figs. 12 and 13 using the same spatial and temporal scales and rainfall thresholds as in Figs. 10 and 11. As in the case of the results for reliability, the results here indicate that the forecasts are better in terms of resolution with increasing spatial and temporal scale. At high rainfall thresholds the outperformance of MF2km over MF10km in resolution is obviously represented, which is similar to the verification results at grid scales. However, despite its worse resolution in predicting heavy rainfall in comparison to that of MF2km, MF10km is better than MF2km in predicting light rains in term of resolution. The reason for this can be traced back to the biases of both systems as plotted in Fig. 3 with the

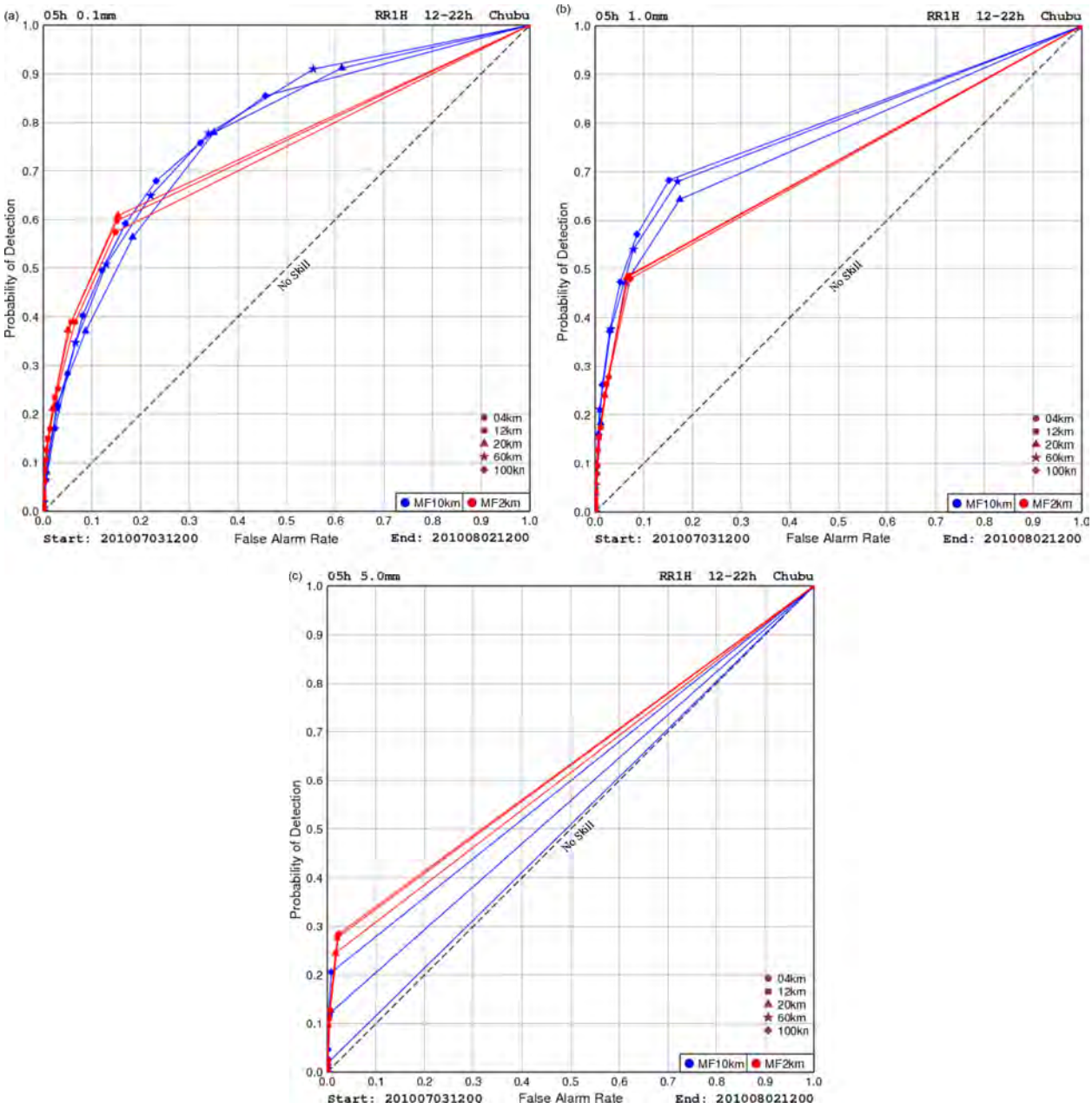

Fig. 12. Same as Fig. 8 but for temporal scale of 5 hours and spatial scales of 4, 12, 20, 60 and 100km.

control forecasts acting as the representatives. It is the underestimation of intense rains of MF10km forecasts that accounts for the superiority of MF2km over MF10km both in reliability and resolution. In the same manner, the same property, but of light rains, of MF2km forecast explains why MF10km outperforms MF2km at light rainfall thresholds.

After examining two systems in terms of reliability and resolution, the performances of two systems are now summarised with the FSS extended in time and ensemble space, a procedure that is analogous with the use of BSS in the traditional verification at grid scales. This summarised evaluation is quantified in Fig. 14 where the ensemble FSSs from MF10km and MF2km under the form of extended intensity-scale diagrams are depicted. To make the comparison between two systems easy, Fig. 14 also plots the differences between the ensemble FSSs of MF2km and MF10km, which were computed as the subtraction of MF2km FSSs by MF10km FSSs. Note that this was done only for the spatial scales resolved by both systems (larger than 20 km).

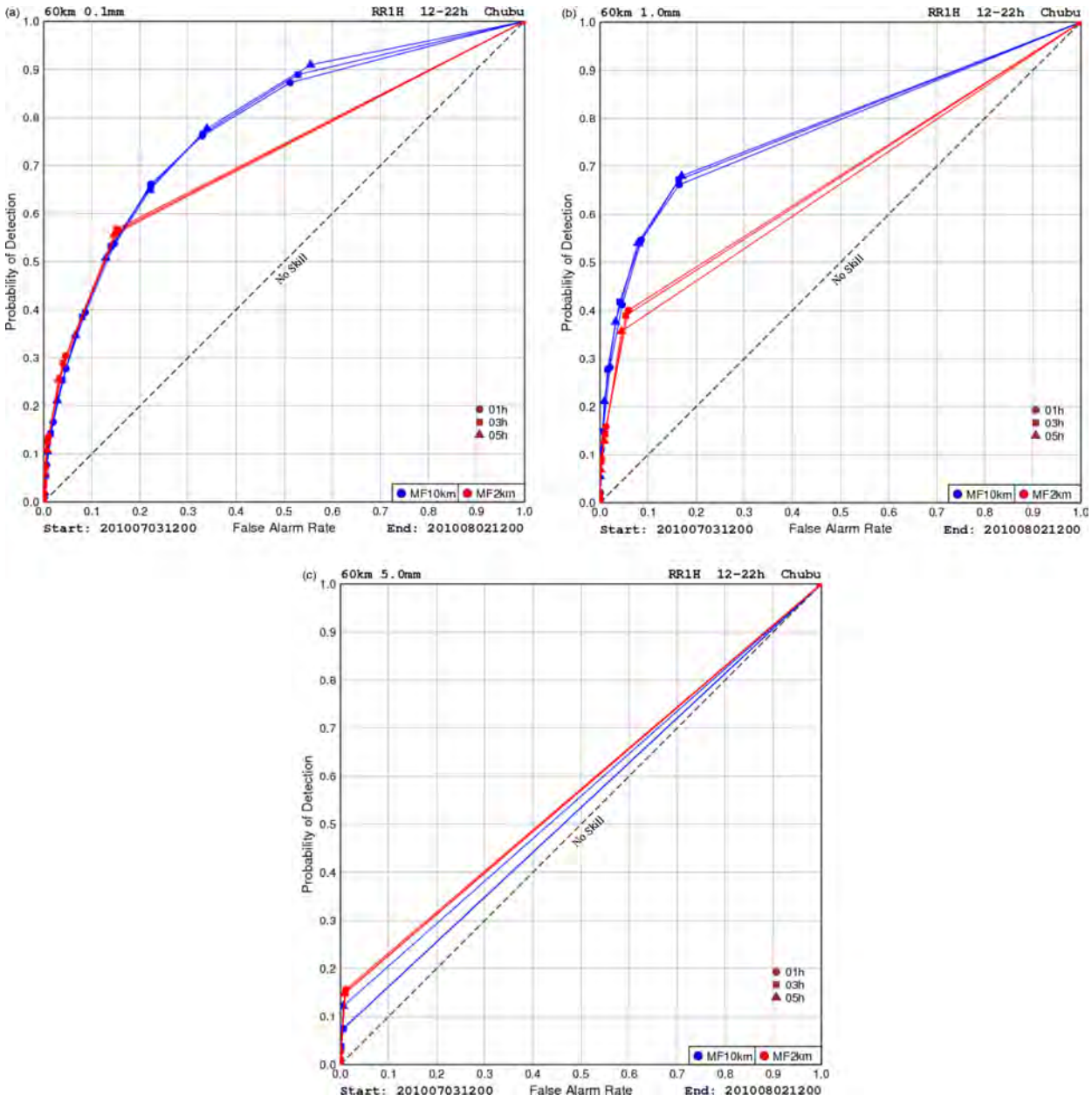

Fig. 13. Same as Fig. 8 but for spatial scale of 60 km and temporal scales from 1 to 5 hours.

Two distinct changes can be identified in the intensity-scale diagram of MF10km. The first change occurs between the 2 and 5 mm h^{-1} rainfall thresholds where FSSs drop sharply. This implies that the performance of MF10km decreases rapidly when rainfall thresholds become close to 5 mm h^{-1}. The performance is lost at the second change between the 5 and 10 mm h^{-1} rainfall thresholds, where the FSS values are smaller than 0.4 for every possible spatial and temporal scale combination. Such distinct changes are not found in the intensity-scale diagram of MF2km. The FSSs in this diagram vary smoothly with rainfall threshold and show a certain skill at high rainfall thresholds if appropriate spatial and temporal scales are considered.

Inter-comparison between two systems in predicting hourly precipitation was performed using the intensity-scale diagram for the FSS differences (the diagram in the bottom of Fig. 14). The most remarkable thing that can be identified in this figure is the large positive FSS differences at high rainfall thresholds. This assesses the outperformance of MF2km over MF10km with respect to heavy rains (greater than or equal to 5 mm h^{-1} in the intensity-scale diagram). This fact is very different from the

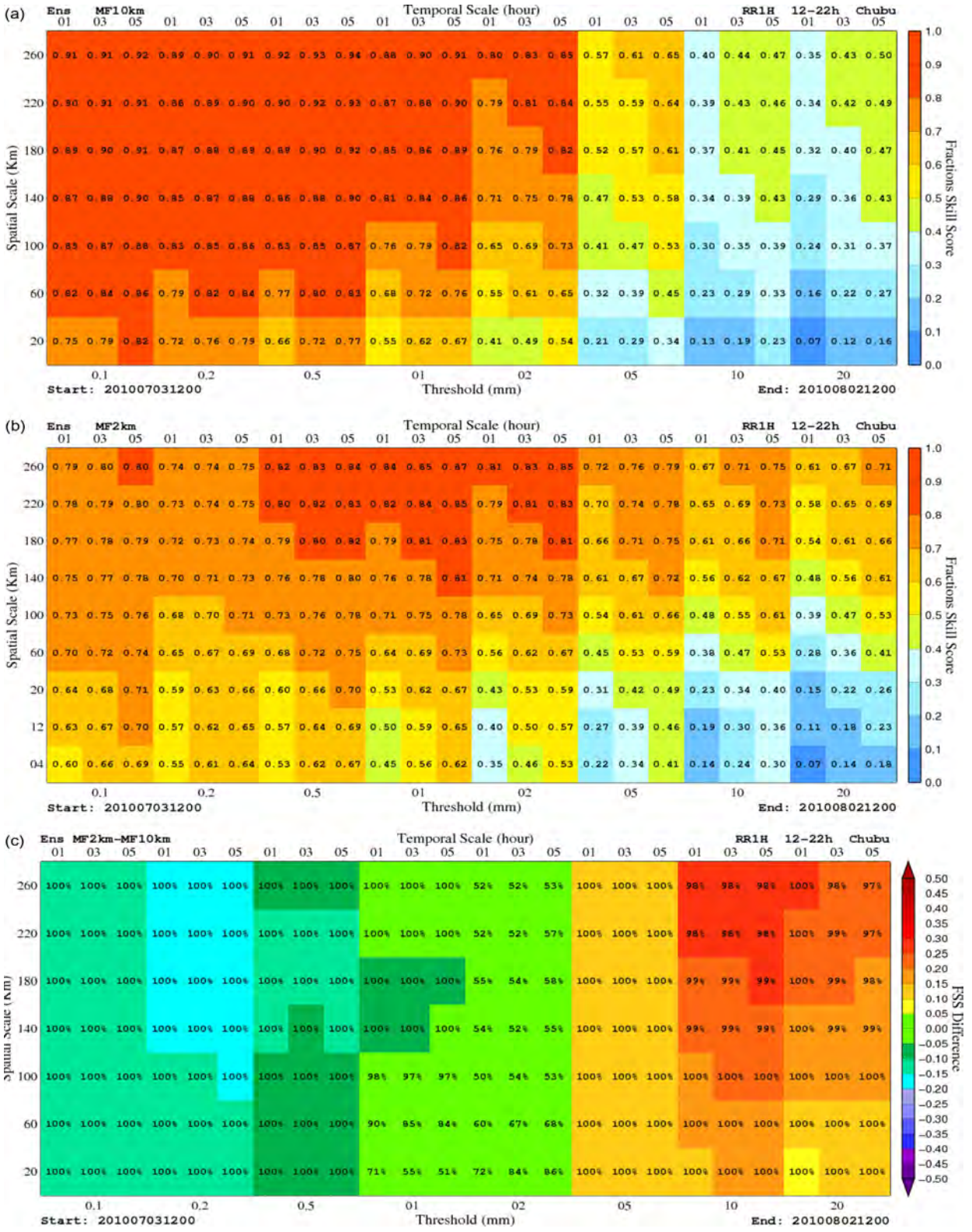

Fig. 14. Extended intensity-scale diagrams showing FSSs from MF10km (top) and MF2km (centre) hourly precipitation forecasts and their differences (bottom). The small number inside each box indicates the confidence level saying the difference value is significantly greater than zero.

implication of the traditional probabilistic verification shown in Fig. 9, where the BSS scores of MF2km are inferior to those of MF10km in the very intense rain regime. In contrast, at rainfall thresholds less than 0.5 mm h^{-1}, the FSS differences have negative values, implying that MF10km forecasts outperform MF2km forecasts if all types of rain are assumed as light rains. For the remaining thresholds, the differences are slightly small with both negative and positive values. A statistical test was carried out to assess whether these small differences are significant or not. Due to its simplicity and robustness, the block bootstrap method was used with 15 000 samples. The assessment that the differences between the FSS values of MF10km and MF2km are not artefacts of computing is represented in the intensity-scale diagram as confident levels in percentages. These confident levels say that those small differences are insignificant and both systems have the same performance at medium rainfall thresholds. This statistical test also reconfirms the outperformance of MF2km over MF10km in predicting heavy rain and the outperformance of MF10km over MF2km in predicting light rains with confident levels of 100%.

So far, as the verification results show, MF10km is more reliable than MF2km in predicting light rains and MF2km more reliable in predicting moderate and heavy rains in contrast. In terms of resolution, MF10km is better than MF2km with respect to light rains and worse than MF2km with respect to heavy rains. The extended FSSs have summarised all those results in a remarkable way. This new insight into the model performances is one of the advantages of FSS in comparison with the traditional BSS.

6. Summary and concluding remarks

In this study, the FSSs verification method extended in time and ensemble space has been presented in order to investigate the value of high-resolution EPS on short-time precipitation forecast. This extension was done by incorporating the time dimension and the ensemble dimension in defining observation and forecast fractions. The mathematical treatment is similar to the original one when the fractions defined in two-dimensional space were now redefined in four-dimensional space. Due to the impact of small-scale processes on verification of short-time precipitation, for example, hourly precipitation, accounting for this source of uncertainties is important. Although ensemble forecast is used as a method to take into account small-scale variability, the problem associated with uncertainties from small-scale processes in high-resolution ensemble verification remains. Combination of ensemble forecast with neighbourhood idea will compensate for this shortcoming.

The new method was tested with the forecast dataset of two ensemble systems MF10km and MF2km of the resolution of 10 and 2 km, respectively, in July 2010. The behaviours of FSS when including the time dimension and the ensemble dimension were examined separately. To explore the relationship between spatial and temporal scales, the intensity-scale diagram was redesigned, allowing both spatial and temporal scales to be displayed in one diagram. The new intensity-scale diagram revealed that by adding temporal scales, FSSs at small spatial scales have similar values as FSSs at large spatial scales without considering temporal scales, which is important if the forecast concerns small scales. The experiment with FSS in ensemble space highlighted the ensemble FSS, which is computed from all ensemble members, as a representative for the FSS of ensemble forecast.

The extended FSS was further applied in verification of MF10km and MF2km forecasts. As the first step, verification based on the traditional scores was performed in terms of reliability and resolution. The BSSs indicates that both systems do not have skill with respect to moderate and heavy rains, while the subjective evaluation on the accumulated rainfalls of the control forecasts suggests a different view. In the next step, the neighbourhood concept was introduced into the traditional verification methods for reliability and resolution. With the change of view from grid scales to larger spatial and temporal scales, both reliability and resolution of two systems increase with increasing spatial or temporal scale. In terms of reliability, MF2km is more reliable than MF10km in predicting moderate and heavy rains. This assessment is not true in predicting light rains. In terms of resolution, MF10km has a better resolution than MF2km in predicting light and moderate rains. However, this reverses in predicting heavy rains when MF2km is better considerably. The reliability and resolution with the spatial–temporal scale of 60 km and 1 hour was almost the same as that with the spatial–temporal scale of 20 km and 5 hours. This result suggests that the ratio of equivalent scales in space and time in the fractions verification (10 km h^{-1}) is affected by the spatiotemporal scales of the meso-scale phenomena. Further investigations should be made to confirm this implication.

Above assessments are reproduced in a compact way by using the extended FSS. MF2km clearly outperform MF10km with respect to heavy rains. In contrast, MF10km is slightly better than MF2km with respect to light rains. This result suggests that the horizontal resolution of 2 km is not necessarily fine enough to completely remove the convective parameterisation.

We used perturbations from JMA's 1-week global EPS for initial and lateral boundary perturbations in our meso-scale EPSs. This method is simple but not necessarily best

for initial perturbations as demonstrated by Saito et al. (2011b). Test of a cloud resolving ensemble prediction using a local ensemble transform Kalman filter is underway at MRI (e.g., Seko et al., 2011), and the validation of its QPF performance is our future subject.

References

Ahijevych, D., Gilleland, E., Brown, B. G. and Ebert, E. 2009. Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*. **24**, 1485–1497.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare, S. E. 2008. The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc*. **134**, 703–722.

Casati, B., Ross, G. and Stephenson, D. B. 2004. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl*. **11**, 141–154.

Clark, A. J., Kain, J. S., Stensrud, D. J., Xue, M., Kong, F. and co-authors. 2011. Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev*. **139**, 1410–1418.

Davis, C. A., Brown, B. G. and Bullock, R. G. 2006. Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. *Mon. Wea. Rev*. **134**, 1772–1784.

Duan, Y., Gong, J., Du, J., Charron, M., Chen, J. and co-authors. 2012. An overview of the Beijing 2008 Olympics Research and Development Project (B08RDP). *Bull. Amer. Meteor. Soc*. **93**, 381–403. DOI: 10.1175/BAMS-D-11-00115.1.

Ebert, E. 2008. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteor. Appl*. **15**, 51–64.

Ebert, E. and McBride, J. L. 2000. Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrol*. **239**, 179–202.

Gallus, W. A. 2010. Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*. **25**, 144–158.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B. and Ebert, E. 2009. Intercomparison of spatial forecast verification methods. *Wea. Forecasting*. **24**, 1416–1430.

Gaudet, B. and Cotton, W. R. 1998. Statistical characteristics of a real-time precipitation forecasting model. *Wea. Forecasting*. **13**, 966–982.

Hirahara, Y., Ishida, J. and Ishimizu, T. 2011. Trial operation of the local forecast model at JMA. *CAS/JSC WGNE Res. Act. Atmos. Ocea. Model*. **41**, 5.11–5.12.

Honda, Y. and Sawada, K. 2008. A new 4D-Var for mesoscale analysis at the Japan meteorological agencyva. *CAS/JSC WGNE Res. Act. Atmos. Ocea. Model*. **38**, 01.7–01.8.

Kunii, M., Saito, K., Seko, H., Hara, M., Hara, T. and co-authors. 2011. Verification and intercomparison of mesoscale ensemble prediction systems in the Beijing 2008 Olympics Research and Development Project. *Tellus*. **63A**, 531–549.

Lack, S., Limpert, G. L. and Fox, N. I. 2010. An object-oriented multiscale verification scheme. *Wea. Forecasting*. **25**, 79–92.

Marsigli, C., Boccanera, F., Montani, A. and Paccagnella, T. 2005. The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification, *Nonlin. Proc. Geophys*. **12**, 527–536.

Marsigli, C., Montani, A. and Paccagnella, T. 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteor. Appl*. **15**, 125–143.

Mittermaier, M. 2007. Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Quart. J. Roy. Meteor. Soc*. **133**, 1487–1500.

Mittermaier, M. and Roberts, N. 2010. Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*. **25**, 343–354.

Nagata, K. 2011. Quantitative precipitation estimation and quantitative precipitation forecasting by the Japan meteorological agency. *RSMC Tokyo–Typhoon Center Technical Review*. **13**, 37–50. Online at: http://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/techrev/text13-2.pdf.

Roberts, N. M. and Lean, H. W. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*. **136**, 78–97.

Saito, K. 2012. The Japan Meteorological Agency nonhydrostatic model and its application to operation and research. In *Atmospheric Model Applications*. InTech, 296 pp.

Saito, K., Fujita, T., Yamada, Y., Ishida, J., Kumagai, Y. and co-authors. 2006. The operational JMA nonhydrostatic meso-scale model. *Mon. Wea. Rev*. **134**, 1266–1298.

Saito, K., Seko H., Kawabata, T., Shoji, Y., Kuroda, T. and co-authors. 2011a. Studies at MRI toward cloud resolving ensemble NWP. *Presentation at 11th EMS Annual Meeting*. Online at: http://presentations.copernicus.org/EMS2011-527_presentation.pdf.

Saito, K., Hara, M., Kunii, M., Seko, H. and Yamaguchi, M. 2011b. Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the meteorological research institute for the WWRP Beijing 2008 olympics research and development project (B08RDP). *Tellus*. **63A**, 445–467.

Saito, K., Seko, H., Kunii, M. and Miyoshi, T. 2012. Effect of lateral boundary perturbations on the breeding method and the local ensemble transform Kalman filter for mesoscale ensemble prediction. *Tellus A*. **64**, 11594, DOI: 10.3402/tellusa.v64i0.11594.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R. and co-authors. 2010. Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*. **25**, 263–280.

Seko, H., Miyoshi, T., Shoji, Y. and Saito, K. 2011. A data assimilation experiment of PWV using the LETKF system–Intense rainfall event on 28 July 2008. *Tellus*. **63A**, 402–414.

Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F. and co-authors. 2011. The Central European limited-area

ensemble forecasting system: ALADIN-LAEF. *Quart. J. Roy. Meteor. Soc.* **137**, 483–502.

Wernli, H., Paulat, M., Hagen, M. and Frei, C. 2008. SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.* **136**, 4470–4487.

Weusthoff, T., Felix, A., Marco, A. and Mathias, W. R. 2010. Assessing the benefits of convection-permitting models by neighborhood verification: examples from MAP D-PHASE. *Mon. Wea. Rev.* **138**, 3418–3433.

Zepeda-Arce, J., Foufoula-Georgiou, E. and Droegemeier, K. K. 2000. Spacetime rainfall organization and its role in validating quantitative precipitation forecastsm. *J. Geophys. Res.* **105**, 10129–10146.